GigaScience, 2024, **13**, 1–13
DOI: 10.1093/gigascience/giae071
RESEARCH

MISIP: a data standard for the reuse and reproducibility of any stable isotope probing-derived nucleic acid sequence and experiment

Abigayle Simpson[1], Elisha M. Wood-Charlson[2,‡], Montana Smith[3,‡], Benjamin J. Koch[4], Kathleen Beilsmith[5], Jeffrey A. Kimbrel[6], Matthew Kellom[7], Christopher I. Hunter[8], Ramona L. Walls[9], Lynn M. Schriml[10], and Roland C. Wilhelm[1,*]

[1]Department of Agronomy, Lilly Hall of Life Sciences, Purdue University, West Lafayette, IN 47907, USA
[2]Environmental Genomics and Systems Biology Division, E.O. Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA 99354, USA
[4]Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, AZ 86011, USA
[5]Data Science and Learning Division, Argonne National Laboratory, Lemont, IL 60439, USA
[6]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA
[7]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[8]GigaScience, BGI-Hong Kong, Hong Kong
[9]Data Collaboration Center, Critical Path Institute, Tucson, AZ 85718, USA
[10]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD 21201, USA
*Correspondence address: Roland C. Wilhelm, Lilly Hall of Life Sciences, 915 West State Street, Purdue University, West Lafayette, IN 47907, USA. E-mail: rcwilhelm@purdue.edu
‡Authors contributed equally.

Abstract

DNA/RNA-stable isotope probing (SIP) is a powerful tool to link *in situ* microbial activity to sequencing data. Every SIP dataset captures distinct information about microbial community metabolism, process rates, and population dynamics, offering valuable insights for a wide range of research questions. Data reuse maximizes the information derived from the labor and resource-intensive SIP approaches. Yet, a review of publicly available SIP sequencing metadata showed that critical information necessary for reproducibility and reuse was often missing. Here, we outline the Minimum Information for any Stable Isotope Probing Sequence (MISIP) according to the Minimum Information for any (x) Sequence (MIxS) framework and include examples of MISIP reporting for common SIP experiments. Our objectives are to expand the capacity of MIxS to accommodate SIP-specific metadata and guide SIP users in metadata collection when planning and reporting an experiment. The MISIP standard requires 5 metadata fields—isotope, isotopolog, isotopolog label, labeling approach, and gradient position—and recommends several fields that represent best practices in acquiring and reporting SIP sequencing data (e.g., gradient density and nucleic acid amount). The standard is intended to be used in concert with other MIxS checklists to comprehensively describe the origin of sequence data, such as for marker genes (MISIP-MIMARKS) or metagenomes (MISIP-MIMS), in combination with metadata required by an environmental extension (e.g., soil). The adoption of the proposed data standard will improve the reuse of any sequence derived from a SIP experiment and, by extension, deepen understanding of *in situ* biogeochemical processes and microbial ecology.

Keywords: stable isotope probing, minimum information standard, MIxS, amplicon, metagenome, metatranscriptome, MIMARKS, MIMS, microbial ecology

Introduction

The invention of DNA/RNA-stable isotope probing (SIP) was a groundbreaking achievement that continues to advance our knowledge of microbiology, microbial ecology, and biogeochemistry [1, 2]. SIP provides a method to link sequencing data with microbial activity resulting from the incorporation of an isotopically labeled compound of interest (*isotopolog*) into the nucleic acids of metabolically active populations (Fig. 1A). Subsequent innovations have improved the utility of SIP for quantifying the differential growth rates of populations within whole microbial communities [3–5]. To achieve this, a typical SIP experiment generates large amounts of sequencing data, given the necessity of sampling multiple density gradient fractions and the use of paired controls (Fig. 1B). Despite fundamental similarities in SIP experiments, there is no consistent vocabulary to catalog the metadata generated. Additionally, the metadata needed to track the composition and handling of nucleic acids during the SIP procedure are frequently absent in sequence archives, impairing the reproducibility of SIP studies and data reuse (Fig. 2A).

The need to facilitate metadata standardization of SIP sequencing data is increasing, as the number of studies generating SIP sequencing data has been rising year upon year (Fig. 2B), with further growth expected due to improvements in automated sample processing [6]. Furthermore, the reuse of SIP sequence data has considerable value given the expense and labor involved in these experiments and the information gained by extrapolating across various isotopologs or study conditions [7–9]. Here, we propose a minimum set of required metadata terms for SIP-derived sequencing data, as well as a recommended set that embodies the best practices in acquiring and reporting SIP sequencing data.

Received: June 1, 2024. **Revised:** July 29, 2024. **Accepted:** August 27, 2024

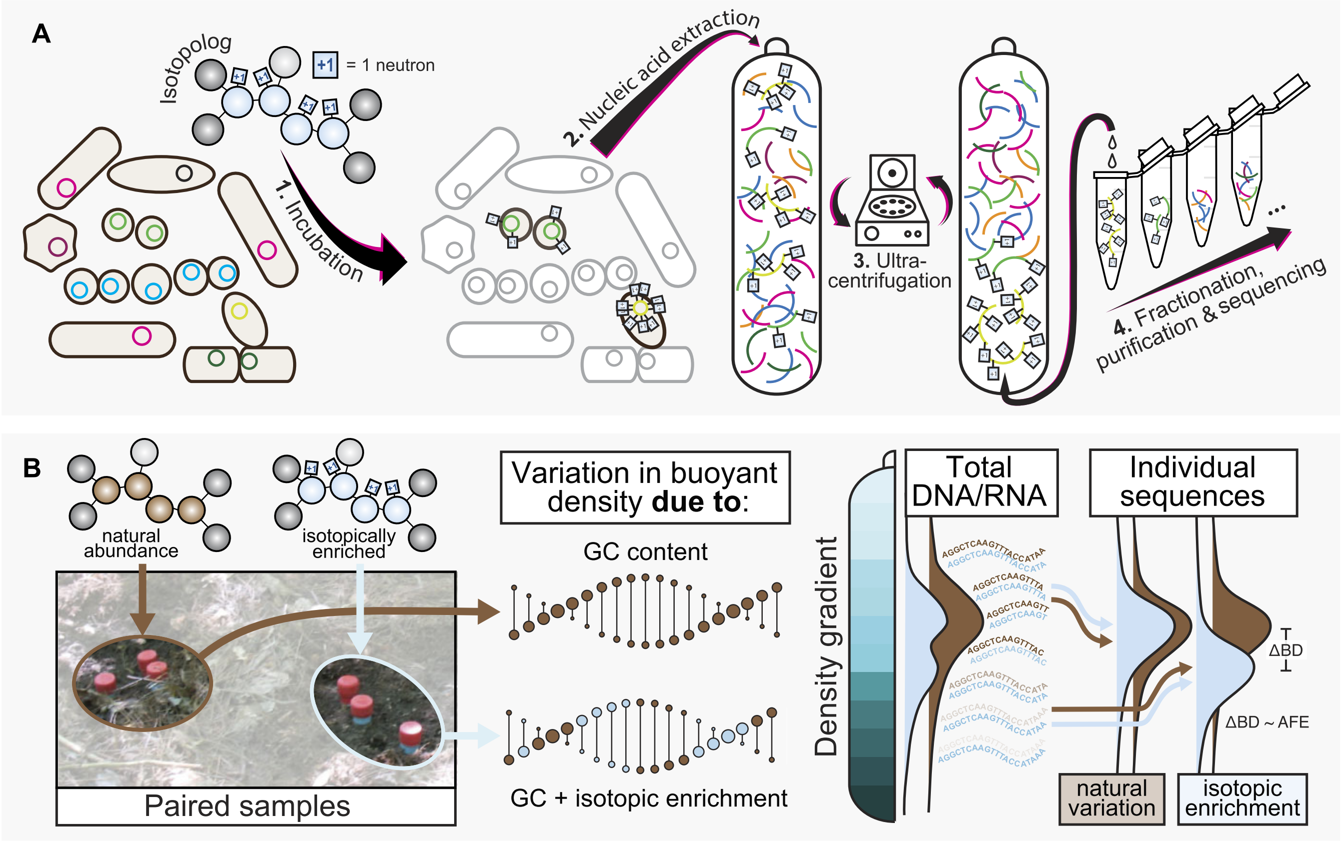

Figure 1: An overview of the key principles of the DNA/RNA-SIP method, illustrating the impact of density gradient separation on sequencing data composition (A) and the need for paired samples supplied with different isotopologs to distinguish natural variation in buoyant density from the effects of isotopic enrichment (B). Panel A shows the standard series of steps performed to isotopically enrich and separate nucleic acids using density gradient ("isopycnic") ultracentrifugation. In this case, an isotopically enriched isotopolog is incubated in the presence of a microbial community, during which the nucleic acids of active cells incorporate artificially high concentrations of the isotope (step 1). Whole nucleic acids are extracted from the sample (step 2) and centrifuged at high force to establish a density gradient and separation of nucleic acids based on buoyant density (step 3). The final step is to fractionate the separated nucleic acids, followed by purification and sequencing (step 4). Panel B depicts the use of paired samples supplied with either natural abundance ("unlabeled") or isotopically enriched isotopologs to isolate the variation in buoyant density of DNA/RNA, due to GC content, from the effects of isotopic enrichment. The excess atom fraction (EAF) of isotopes can be estimated on a per sequence basis using the change in buoyant density (ΔBD) between paired samples [3], demonstrating the importance of paired samples.

These terms were selected based on a comprehensive literature review and development with domain experts in the SIP community.

Essential properties of SIP sequence data

In all cases, sequence data generated from a SIP experiment originate from nucleic acid pools that have been fractionated by isopycnic (or "density gradient") separation [10]. The gradient separates nucleic acids based on differences in buoyant density due to the added mass per nucleic acid from the incorporation of heavy stable isotopes (e.g., ^{13}C, ^{15}N, or ^{18}O) during the metabolism of an isotopically labeled source isotopolog (e.g., $^{13}CO_2$, $^{15}NH_4^+$, or $H_2{}^{18}O$). The fractionated pools of nucleic acids are then sequenced to resolve differences in buoyant densities corresponding to isotopic enrichment. Many fractions may be sequenced to determine fine-scale isotopic enrichment ("density-resolved SIP"), or fractions may be pooled before sequencing to compare coarse differences in buoyant density. Either way, each nucleic acid sample typically generates multiple sequencing libraries (Fig. 1A).

There is natural variation in the buoyant density of nucleic acids due to the effect of GC content on genome density [11]. This variability creates sample-specific buoyant density distributions of nucleic acids based on the genomic composition of the biological community under study. To control for this, the standard SIP approach involves comparing sequence data generated from identically treated sample pairs: one that received an unlabeled isotopolog (i.e., natural abundance) and another that received an artificially labeled isotopolog [10] (Fig. 1B). Alternative SIP approaches without paired controls are possible based on the modeling of expected natural abundance distribution patterns from sequence data [12]. In either case, the reproducibility and reuse of SIP sequence data cannot be achieved without information about the position from which the nucleic acids originated in the density gradient and/or its corresponding paired control, as well as information about the stable isotope(s) and source isotopolog compound(s) used.

The case for a SIP-specific data standard

At present, there is no convention for the handling of SIP sequence metadata despite the archival of hundreds of datasets in public databases, spanning over 2 decades of sequencing types (clone libraries to shotgun metagenomes) from diverse environments (Supplementary Fig. S1). A formal standard describing the minimum information for any SIP sequence is needed to ensure the adoption of FAIR principles (Findability, Accessibility, Interoperability, and Reusability) for data reuse [13]. Currently, there are

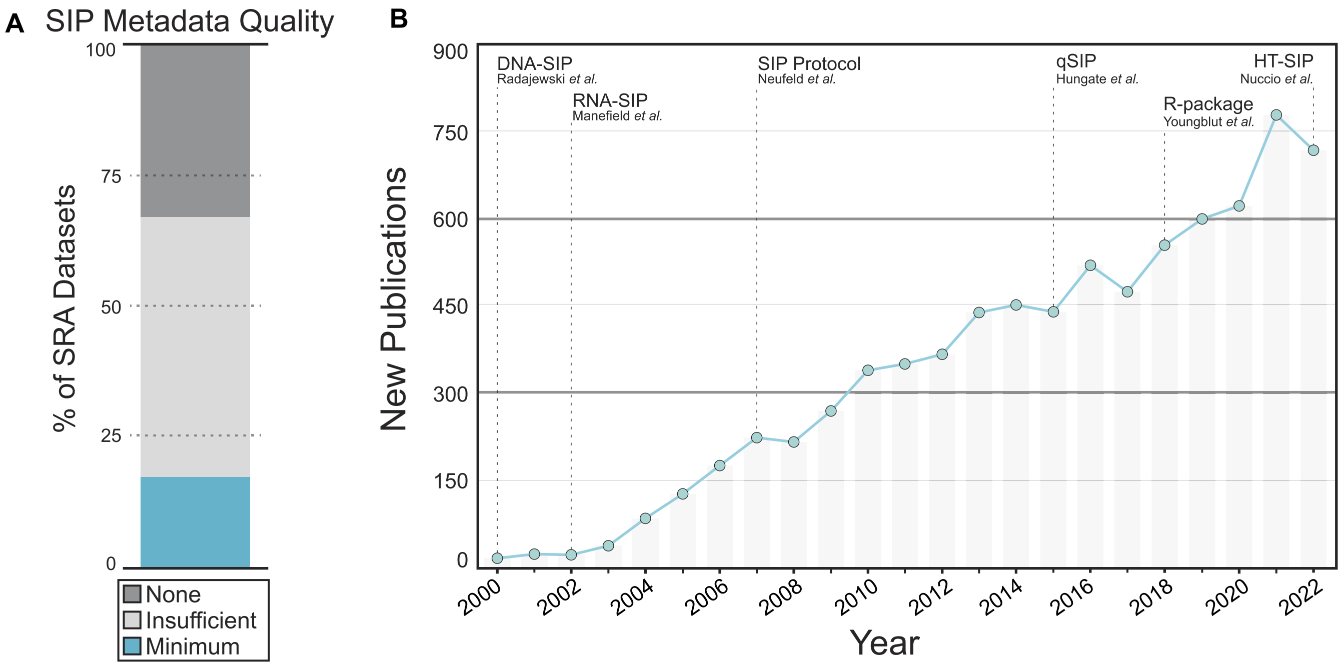

Figure 2: A summary of the quality of metadata associated with SIP sequencing data available in the Sequence Read Archive (A) and the growth in the number of DNA/RNA-SIP studies published over time (B). In (A), an overview of the metadata quality of studies with SRAs containing more than 5 samples. SRAs that reported the isotopolog, isotopolog label status, and gradient position for each entry met the "Minimum" requirements, while those that reported at least one of these items were "Insufficient," and those that reported none were categorized as "None." In (B), the new number of published studies by year was identified using the search term: "Stable Isotope Probing" and "DNA" or "RNA" in Google Scholar. A timeline showing the major advances in the SIP methodology was included.

no accommodations for essential SIP metadata or a stable identifier for SIP sequencing projects, requiring users to glean this information from the study description or associated publication, which poses challenges to findability. The absence of data labeling requirements creates the risk of naive users misinterpreting ambiguously labeled SIP data and failing to account for biases caused by density gradient fractionation. Furthermore, there is no common vocabulary for the diverse types of SIP sequence metadata generated during the course of an experiment. The formalization of a specific, richly described, and consistent vocabulary is critical for interoperability among SIP sequence data, facilitating comparisons across SIP studies. The creation of SIP-specific metadata fields will greatly improve the machine readability, analyses, and interpretation of SIP sequence data.

For these reasons, we propose the Minimum Information for any Stable Isotope Probing Sequence (MISIP) data standard designed to capture critical information about the origin of sequences from SIP experiments. MISIP contains SIP-specific fields in combination with other Minimum Information for any (x) Sequence (MIxS) standards, including the "Minimum Information about a MARKer gene Sequence" (MISIP-MIMARKS) [14] and the "Minimum Information about a Metagenome Sequence" (MISIP-MIMS) [15] checklists, which have been developed in collaboration with the Genomic Standards Consortium (GSC) [16]. MISIP must also be paired with a MIxS extension to describe the environmental context from which nucleic acids were extracted (e.g., water, soil, or host associated). MISIP was developed after an extensive review of existing SIP literature (Supplementary Table S1), conversations with the SIP research community, solicitations for comment on a preprint [17], and via a survey of active SIP research groups. The standard then underwent iterative refinement with the Compliance and Interoperability Working Group (CIG) of the GSC. All proposed field names are unique and nonredundant according to queries of all current MIxS checklists and extensions described by the GSC. When possible, terms are compliant with standard SIP terminology [18], with labels of required or recommended to enable machine-actionable validation.

The minimum information for any SIP-derived sequence

The MISIP standard includes a required and a recommended set of metadata fields (Table 1). Required fields are essential for documenting the fundamental changes in nucleic acid composition due to the SIP method, without which archived SIP sequence data risk being misused and losing their scientific utility. Recommended fields capture information that most SIP practitioners agree is vital for robust validation and meaningful analyses, including quantitative SIP analyses and cross-study comparisons. In the following sections, we provide descriptions of the metadata fields in the MISIP standard and justify their designation as required or recommended in the associated sequencing depositor checklist. We intend this information to serve as a reference for users of MISIP and to help newcomers to SIP methods to collect and curate better metadata. We have also provided examples of SIP metadata curated according to the MISIP standard, including examples of a common gradient pooling approach (Supplementary Table S2), a more complex SIP experiment (Supplementary Table S3), quantitative SIP (qSIP; Supplementary Table S4), and examples of curation data derived from ^{18}O (Supplementary Table S4) and ^{15}N experiments (Supplementary Table S5), along with several other cases (Supplementary Tables S5–S9). MISIP will be maintained as a living standard by the GSC, ensuring it may be adapted to serve the evolving methods and needs of the SIP community.

Table 1: The minimum set of SIP-specific metadata required or recommended by the MISIP data standard. The table summarizes the standardized vocabulary, data format, and whether it is required or recommended (i.e., optional). Descriptions of each metadata field and justifications for their inclusion are provided in the text below.

Slot	Term	Format	Criteria
isotope	Isotope(s)	Element with atomic mass (e.g., ^{13}C, ^{15}N, ^{2}H, or ^{18}O)	Required
isotopolog*	Isotopolog (isotope source/substrate)	PubChem Compound Identification (CID) number or, if an undefined mixture, “0”	Required
isotopolog_label	Isotopolog label status	Defined category (“isotopically labeled” or “natural abundance”)	Required
isotopolog_approach	Labeling approach (number of labeled isotopologs supplied)	Defined category (“single” or “multiple”)	Required
gradient_position	Gradient position	An integer designating the gradient position from heaviest (= 1) to lightest	Required
gradient_pos_density	Density of gradient position	Density of the fraction in g/mL	Recommended†
gradient_pos_rel_amt	Relative amount of DNA in the gradient position	Proportion of total DNA (min = 0, max = 1)	Recommended†
sip_method	Method for fractionating	DOI	Recommended†
source_mat_id	ID of sample prior to fractionating	Any text	Recommended
isotopolog_atom_frac	Atom fraction of isotopolog	Proportion (min = 0, max = 1)	Recommended
isotopolog_atom_pos	Set of labeled atoms in isotopolog	InChI label	Recommended
isotopolog_dose	Dose of isotopolog	Amount of isotopolog in ppm	Recommended
nucleobase_atom_frac	Atom fraction of nucleobases	Proportion (min = 0, max = 1)	Recommended
isotopolog_incu_time	Incubation time	Hours	Recommended
chem_administration	Additional substrates	Any text (comma separated)	Recommended
internal_standard	Internal standard method	DOI	Recommended

*Specify the attributes of the material delivered to the biological system. If the isotopolog is an undefined mixture, follow instructions provided in the text.
†Highly recommended, but not strictly required for metadata archival.

Descriptions of required MISIP fields

Isotope

The specific stable isotope(s) supplied to the biological system is required information, since the stoichiometry of each element in nucleic acids can influence the magnitude of shift in buoyant density of the nucleic acids used to generate SIP sequencing data. For example, fully ^{13}C-labeled DNA would produce a larger increase in buoyant density compared to fully ^{15}N-labeled DNA, owing to the unequal ratio of approximately 5 carbons to 1 nitrogen in DNA. The ***isotope*** field specifies the element and mass number (e.g., ^{18}O recorded as “18O”) of the stable isotope of interest. This field will correspond to the same stable isotope regardless of whether the concentration of the isotope was artificially enriched (e.g., 0.99 atom fraction ^{18}O) or occurred at natural abundance (~0.00205 atom fraction ^{18}O), often referred to as a “control.”

Isotopolog

The central aim of a SIP experiment is to link the isotopic enrichment of nucleic acids to the metabolism of an isotopolog source (or “substrate”). The chemical properties of the isotopolog determine how to interpret the underlying metabolic activity that produced the isotopic enrichment of nucleic acids and the associated changes in the composition of sequence data. For example, certain isotopologs, such as $H_2{}^{18}O$, are used to characterize the whole metabolic activity of a microbial community [19], while others, such as ring-$^{13}C_6$-labeled phenolic acid, are used to target specific metabolic activity [20]. Surprisingly, the number of accessioned SIP experiments that report the isotopolog in the associated metadata is low (<30%; Supplementary Table S1), making it an often overlooked, but essential, attribute of SIP sequencing data. The ***isotopolog*** field specifies the PubChem Compound Identification (CID) number for the isotopolog serving as the isotope source (e.g., 6255 for maltose). If a PubChem CID does not exist for the isotopolog molecule, users should create a new CID using the NCBI PubChem website. If the isotopolog is an undefined chemical mixture, the *isotopolog* field should specify “heterogenous source,” and the additional field, *hetero_isotopolog*, should be used to specify the nature of the isotopolog mixture in a separate column in text form (e.g., Supplementary Table S6). In cases where a sample was not amended with an isotopolog (e.g., an unamended control), the *isotopolog* field should list “none,” and other isotopolog-related fields should specify “not applicable.”

Isotopolog label

The standard SIP method requires the pairing of samples that have been supplied with either an isotopically labeled or natural abundance (“unlabeled”) isotopolog to account for shifts in buoyant density due to variation in the GC content of genomes [10, 11]. The isotopolog label status is essential to determine whether the buoyant density distribution of nucleic acids reflects isotopic labeling or natural variation. The inclusion of paired samples is not required by MISIP, since the gradient distribution of natural abundance nucleic acid fragments can be modeled [12]. However, the ***isotopolog_label*** field is required as it specifies whether the corresponding isotopolog contains the natural abundance (“natural abundance”) or artificially enriched (“isotopically labeled”) concentration of stable isotopes.

In cases where one control sample has been amended with several natural abundance isotopologs and is paired with numerous samples supplied with individual isotopically enriched isotopologs (e.g., [8]), the control sample (*isotopolog_label* = “natural abundance”) should be replicated multiple times in the metadata with each corresponding *isotopolog* field changed to match the paired isotopically labeled source (e.g., Supplementary Table S3).

Isotopolog labeling approach

Several isotopes and/or isotopologs may serve as the source of the isotopic enrichment of nucleic acids in an SIP experiment. For example, researchers can use a dual-labeling approach, as previ-

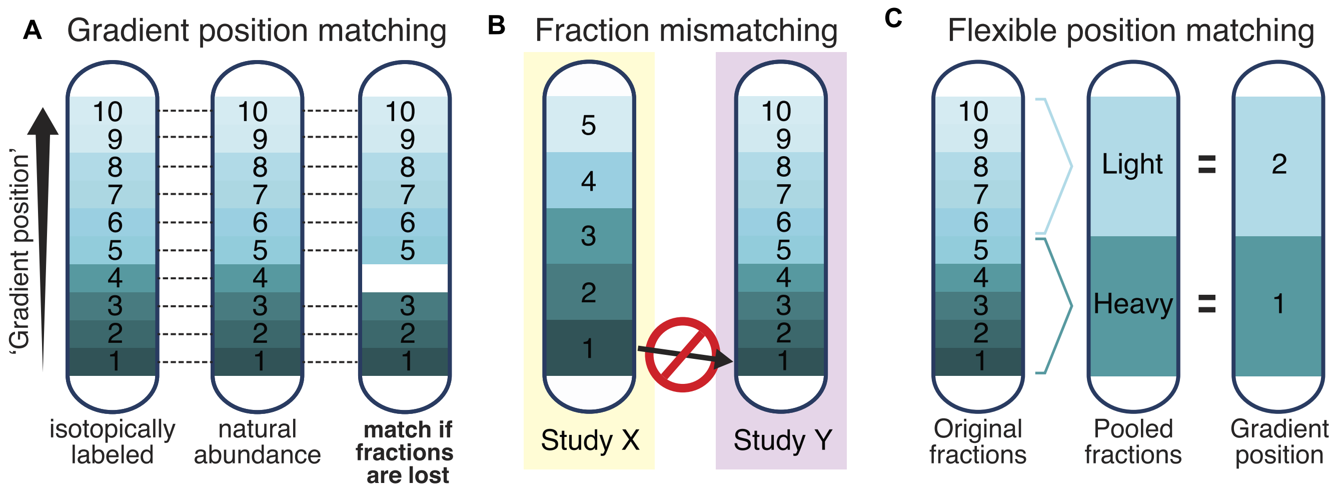

Figure 3: The separation of nucleic acids according to differences in buoyant density due to isotopic enrichment is a defining feature of SIP sequence data. The gradient position designates the location along the density gradient from which the sequenced nucleic acids were recovered. The gradient position generally follows the order in which gradient fractions were collected, accounting for cases where fractions have been lost during processing. Panel A demonstrates the importance of ensuring gradient positions match between paired isotopically labeled and natural abundance samples. Panel B illustrates the disparity in gradient density due to study-specific methods when comparing fraction numbers. Panel C illustrates the use of the gradient position system to accommodate different strategies employed for gradient fraction pooling.

ously performed with $H_2{}^{18}O$ and ^{13}C-glucose [21]. In these cases, multiple isotopes can contribute to shifts in nucleic acid buoyant density, complicating the analysis and interpretation of SIP sequence data. These datasets should be clearly labeled and easily filtered to ensure appropriate data reuse. The ***isotopolog_approach*** field specifies whether the associated SIP experiment utilized a single isotope and isotopolog (*isotopolog_approach* = "single") or multiple isotopes or isotopologs (*isotopolog_approach* = "multiple") within the same sample.

While multiple-labeling approaches make up less than 10% of accessioned SIP studies (Supplementary Table S1), these experiments complicate the reporting of a number of other fields, such as the isotope, isotopolog, and isotopolog-describing fields. In these cases, users should list values for these fields in a consistent order separated with a vertical bar symbol corresponding to each isotope (e.g., *isotope* = ^{13}C | ^{18}O, *isotopolog* = 5793 | 962).

Gradient position

SIP sequencing data are heavily influenced by the differences in the buoyant density of nucleic acids recovered at positions across the density gradient. Thus, MISIP requires information about the gradient position from which the sequenced nucleic acids were recovered. The ***gradient_position*** field is specified as a number starting from the densest ("heaviest") gradient fraction (= 1) moving in sequential order to the least dense fraction ("lightest"), keeping with the order in which a gradient is typically fractionated. MISIP users must take special care to ensure that the gradient position numbers match between paired isotopically labeled and natural abundance samples (Fig. 3A), accounting for any pooling of fractions that may be performed. The inclusion of unfractionated samples, from which fractionated samples were derived, can be denoted with *gradient_position* = –1.

The MISIP standard uses the numerical order of the gradient position to match paired samples because it is the most flexible way to accept SIP sequence data. SIP experiments typically assign a fraction number or measure the buoyant density of the fraction from which nucleic acids were recovered. However, the treatment of fractionated nucleic acid pools will depend on the subjective aims of a SIP experiment, leading to diverse and ad hoc ways in which fractions are pooled prior to sequencing. Defaulting to the numerical order for the position in the density gradient parallels the common approach of assigning a number to each "gradient fraction" obtained. MISIP uses the term "gradient position" to avoid the assumption of equivalence between fractions across studies, since fractionation will yield different volumes and density ranges depending on methodologies (Fig. 3B). When fractions are pooled or sequenced in an ad hoc manner, we strongly urge depositors to provide gradient density measurements (*gradient_pos_density*) to enable normalization among samples and across studies, based on the actual buoyant density range of nucleic acids.

Direct measurement of the buoyant density of each gradient fraction is strongly recommended but not required for several reasons. The measurement of buoyant density, using the refractometric index or by weighing, is not the only method to establish whether nucleic acids have been separated according to differences in buoyant density. Differences in the relative amount of nucleic acids and/or their excess atom fraction can also serve as a measure of density gradient separation [22]. Furthermore, it is possible to use internal standards in lieu of gradient density to obtain a direct measure of the buoyant density distribution of nucleic acids [5]. Consequently, the MISIP standard will populate the MIxS checklist with the gradient density field by default, but acquiring and submitting gradient density data is not required for data archival.

Gradient position is agnostic to the assortment of ways nucleic acids are treated during fractionation. For example, "heavy" and "light" are common designations of the location in the density gradient from which SIP sequence data originate after pooling multiple gradient fractions. In the MISIP gradient position system, these categorical values would be assigned a gradient position of 1 and 2, respectively, which are decoupled from the original fractions numbers (Fig. 3C). The gradient position system has its own pitfalls, with positioning becoming discordant when a fraction is lost or when fractionation is inconsistently initiated, though we anticipate the consistency of fractionation will improve over time due to increasing automation [6].

Depositors must take special care to ensure that, at the very least, the gradient position numbers reflect the gradient distribu-

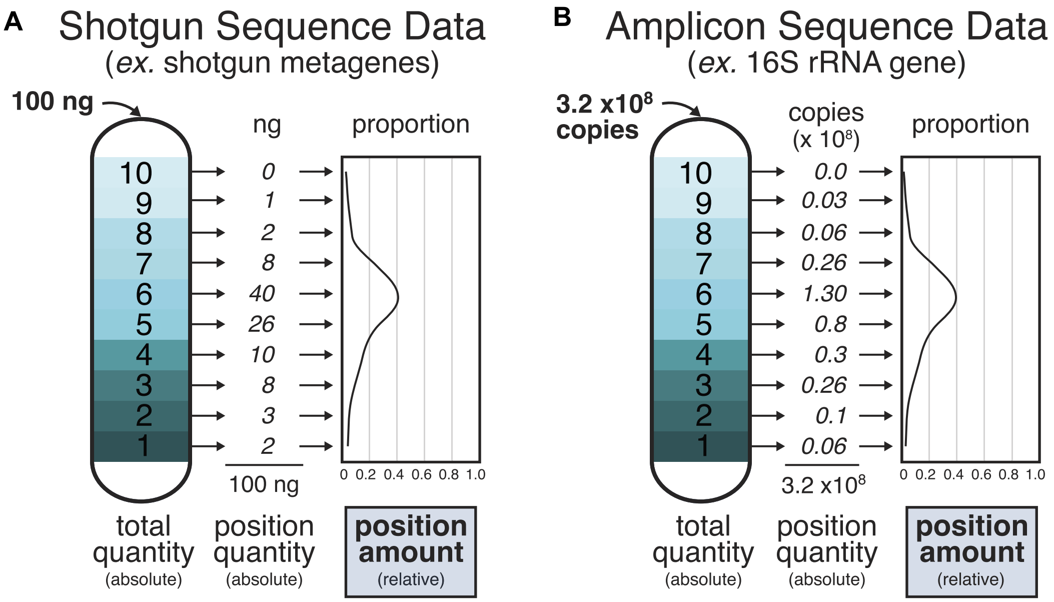

Figure 4: Data on the relative amount of nucleic acid at each gradient position (*gradient_pos_rel_amt*) can be used to calculate taxon-, genome-, or gene-specific EAF using the qSIP method. This value is given as the proportion of nucleic acids at each sequenced gradient position relative to the total nucleic acids available for sequencing. The total amount of nucleic acids depends on the sequencing approach. For shotgun approaches (in A), the *gradient_pos_rel_amt* will correspond to the total mass of nucleic acids (DNA or RNA). For amplicon approaches (in B), the *gradient_pos_rel_amt* will correspond to the total copy number of the targeted gene established via quantitative PCR.

tion between paired isotopically labeled and unlabeled controls, should pairs exist in the dataset. Users of MISIP sequence data must be aware of the potential incongruity between buoyant density and gradient position. These sources of error can be easily resolved by collecting the recommended MISIP metadata, including measurements of gradient density (*gradient_pos_density*), quantifying the distribution of nucleic acids, or using internal standards to evaluate gradient formation [5]. Similarly, if paired samples do not exist, users should provide the information on the method used to validate density gradient formation.

Recommended MISIP sequence metadata

The required set of fields in the MISIP standard were chosen to accommodate the fullest diversity of SIP experimental approaches and sequence data while maintaining the minimum basis for reuse. The minimum required terms should be augmented with the recommended fields to improve FAIRness of submitted metadata and to improve cross-study comparisons and reproducibility [13]. We propose a set of recommended fields and highlight those we consider to be the gold standard (🥇) for supporting the most sophisticated and quantitative reuse of SIP sequencing data. Example metadata that have been curated according to the MISIP standard are available in Supplementary Tables S2 to S9.

Gradient position density 🥇

Measurement of the density of solution recovered at each gradient position is crucial to (i) evaluate the formation of the density gradient, (ii) normalize among gradient fractionated samples, and (iii) calculate the change in buoyant density ("ΔBD") to estimate the degree of isotopic enrichment of sequenced nucleic acids [3]. The ***gradient_pos_density*** field is specified as a numerical value corresponding to the density of gradient solution in grams per milliliter ($g \cdot mL^{-1}$). Measurements of gradient density can be obtained using a refractometer or analytical balance [10]. When measured with a refractometer, ensure the following: (i) ensure that refractometer has been recently calibrated, (ii) use sufficient volume and take rapid measurement to avoid fluctuations caused by evaporation, (iii) operate in a consistent ambient temperature, and (iv) provide information on the methods used to convert between refractive index and gradient density (including any temperature correction) in your SIP methodology (*sip_method*), as previously shown [23, 24].

Relative amount of nucleic acid at gradient position 🥇

Measurement of the relative quantity of sequenced nucleic acids in each density gradient fraction can establish the separation of nucleic acids by buoyant density and is used to estimate the taxon-, genome-, or gene-specific isotopic enrichment, or excess atom fraction (EAF), according to the qSIP method [3, 5]. The ***gradient_pos_rel_amt*** field is specified as the proportion of sequenced nucleic acids relative to the total amount of nucleic acids added to the density gradient prior to ultracentrifugation and must not exceed a value of 1. This measure is also referred to as the "ratio of maximum quantity." For shotgun sequencing data, this proportion could be calculated from total nucleic acids (Fig. 4A), while for amplicon sequence data, this proportion should be calculated from the total copies of the gene target, typically measured by quantitative PCR (Fig. 4B).

SIP methodology 🥇

Density gradient formation and the composition of nucleic acids recovered depend on a range of methodological considerations, including rotor type, run speed, run length, gradient medium (e.g., cesium chloride for DNA or trifluoroacetic acid for RNA), fraction volume, and pooling strategy. These methodological details do not fit cleanly into a data standard. However, this information is vital for data interpretation and to explain potential differences among studies during a meta-analysis. The ***sip_method*** field specifies a DOI corresponding to a protocol, article, or data accession in which the complete methodological details have been provided, including any modifications to standard approaches. This type of information can be stored on Protocols.io with a stable DOI [25] when an alternate DOI is not available at the time of archival.

Source material identity

SIP sequence data are often complicated by the generation of multiple sequencing products from a single nucleic acid extract ("sample"). We recommend that depositors specify a unique identifier that links all postfractionation sequence data to the original (unfractionated) nucleic acid extract. The *source_mat_ID* (MIXS:0000026) is an existing field in the MIxS framework used to indicate the source sample from which nucleic acids were derived [14]. In MISIP, the ***source_mat_ID*** field is specified as a character string that denotes the unique sample identity for the unfractionated nucleic acid source from which all downstream sequence data originate.

Following the MIxS *source_mat_ID* requirements, we urge users to employ a unique Globally Unique IDentifier (GUID) to maintain the link between the origin of a sample and all downstream measurements. We recommended that the unique GUIDs refer to the experimental sample (e.g., soil) or the unfractionated nucleic acid source. Options for resolvable GUIDs include an International Generic Sample Number (IGSN) [26], a BioSample accession number, or an Archival Resource Key, among others. Notably, if the unfractionated source material was sequenced, the sample should be specified by setting *gradient_position* = –1.

Isotopolog atom fraction

The isotopolog atom fraction refers to the average proportion of an isotope in an isotopolog. For example, the atom fraction of isotopically labeled acetate containing an average of one ^{13}C atom per molecule (^{13}C-CH$_3$O$_2$) would be 0.5 atom fraction ^{13}C. The atom fraction of an isotopolog will affect the kinetics of the isotopic labeling of nucleic acids. For example, a low atom fraction will tend to yield more marginal enrichment of nucleic acids and smaller shifts in buoyant density, which can impact the interpretation of sequence data analysis. The ***isotopolog_atom_frac*** field is specified as the atom fraction isotope (as a decimal) of the isotopolog source substrate. Note that this is an average of all isotopolog molecules in the prepared isotopolog source material and is often stated as atom % by the commercial provider of isotopically enriched material. Finally, if the isotopolog was produced in-house (e.g., bacterial cellulose from 0.99 atom fraction ^{13}C-glucose), ensure that the *isotopolog_atom_frac* corresponds with the atom fraction of the final isotopolog used to label nucleic acids (e.g., ^{13}C-labeled bacterial cellulose), not the upstream isotopolog used to generate the labeled substrate (e.g., ^{13}C-glucose). This will require a researcher to measure the isotopic concentration of their isotopolog.

Isotopolog atom position

In cases where an isotopolog is not uniformly isotopically labeled, differences in the molecular position of isotope atoms can alter the proportion of isotope metabolized into nucleic acids. For example, organisms that preferentially metabolize a functional group, or sidechain, might receive more isotopic label than those that metabolize the whole, or parts, of a partially labeled isotopolog [27]. The ***isotopolog_atom_pos*** field is specified as the International Chemical Identifier (InChI) label [28], which designates the set of all isotopically enriched atoms present in the isotopomer [18] of the isotopolog supplied, according to their molecular position. The *isotopolog_atom_pos* should specify the orientation of isotopes in the main isotopomer of the isotopolog source. If the isotopolog consists of more than 1 defined isotopomer, users should list separate InChI labels for each delimited with a vertical bar symbol ("|"). If more than 1 isotopolog has been used, list the *isotopolog_atom_pos* of each isotopolog delimited with a vertical bar in an order that matches the list in the *isotopolog* field. A protocol for generating an InChI label using the InChI open-source chemical structure representation algorithm is available on Protocols.io [29]. The *isotopolog_atom_pos* should only be provided if the isotopolog is a defined compound (i.e., **isotopolog** != "none" or "not applicable").

Isotopolog dose

The concentration of isotopolog ("dose") added to the system will influence the rate and degree of isotopic labeling of nucleic acids. The dose is the mass of isotopolog added per volume of the relevant environmental matrix (e.g., glucose/total volume soil; CH$_4$/total container volume). The dose should reflect the concentration of isotopolog in the biological system, accounting for the dilution by the environmental matrix, not the concentration of the added isotopolog solution (Fig. 5). The dose should also reflect the total cumulative isotopolog added to the system prior to nucleic acid extraction, accounting for multiple additions to the system across time. In cases where the isotopolog is not homogenized within the environmental matrix, the dose should be an estimate of the isotopolog concentration in the sample used for nucleic acid extraction. When estimates are too uncertain or when the concentration of isotopolog in the system is unknown (e.g., root exudates), no dose should be specified, but contextualizing information should be provided in the reference provided in the *sip_method* field. The ***isotopolog_dose*** field is specified as the final concentration of isotopolog added to the system in parts per million (ppm).

Nucleobase atom fraction

Bulk measurement of the isotope content of nucleic acids can be used to assess the rate of isotopic labeling, measure the buoyant density separation of nucleic acids, or validate buoyant density shift-based estimates of EAF. The ***nucleobase_atom_frac*** field is specified as the atom fraction (as a decimal between 0 and 1) of isotope in the nucleic acids pool used to generate sequencing data, as previously described [22].

Incubation time

The isotopic enrichment of nucleic acids depends on the kinetics of isotopolog metabolism and the fluxes of stable isotope in the experimental system. Over time, isotopic labeling of secondary populations will occur due to access to isotopolog-derived metabolites and biomass. The dynamics of cross-feeding of isotopolog-derived biomolecules will influence SIP sequence composition over time. The inclusion of incubation time is recommended to support the calculation of labeling rates (especially when combined with *nucleobase_atom_frac*) and to estimate growth rates using qSIP [30, 31]. The ***isotopolog_incu_time*** field is a numerical field specified as the time in hours (h) from the addition of isotopolog to the end of the incubation period.

Additional substrates

SIP experiments may involve the co-amendment of unlabeled compounds with an isotopolog to serve as growth substrates [32] or other chemical treatments, including inhibitors of specific populations or metabolisms [33]. The chemical modification of growth conditions exerts a profound influence over community

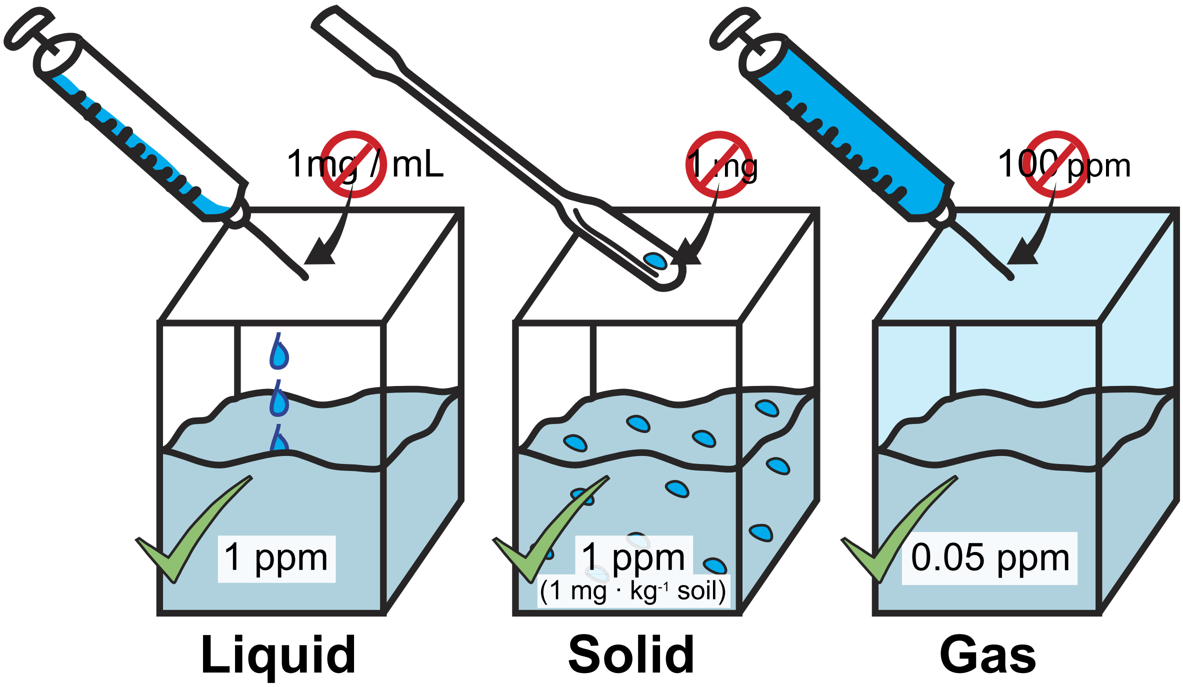

Figure 5: The ***isotopolog_dose*** field provides the concentration, in ppm, of isotopolog in the biological system under study. This field corresponds to the isotopolog exposure of the biological community as measured by the total concentration throughout the system (indicated by the blue fill color) and not the concentration of the amended gas, liquid, or solid material (indicated by the "no symbol"). The numbers provided here are intended to serve as examples.

metabolism and the isotopic labeling of nucleic acids. In these cases, we encourage the inclusion of the ***chem_administration*** field (MIXS:0000751) to specify a comma-separated list of chemical compounds coadministered to a host or environment along with an isotopolog(s).

Internal standard

Internal nucleic acid standards can be used to evaluate density gradient formation and nucleic acid separation, as well as to normalize gradient position across samples. Internal standards are pools of nucleic acids chosen to represent a range in buoyant densities based on characteristics of atom fraction isotope and GC content. For example, pre-centrifugation internal spike-in standards were used to identify samples with anomalous density gradient formation [5]. At present, there is no standard methodology for generating or implementing SIP internal standards. Thus, the ***internal_standard*** field specifies a DOI corresponding to a protocol that provides the methodological details, including the sequence composition, isotopic enrichment, and expected buoyant density distribution, of any internal standard added to sequencing libraries. This information can be provided on Protocols.io and accessioned with a stable DOI [25]. The internal standard DOI may be the same as the SIP methodology DOI so long as the SIP methodology fully describes the characteristics of the internal standards.

Accommodating the complexity of SIP experiments

The MISIP data standard specifies the essential information needed to reliably discern the influence of isotopic labeling on the nucleic acid composition of SIP sequence data (Fig. 6). These advancements, along with recommendations to guide the collection of other valuable metadata, are sufficient for the majority of SIP experiments. However, the diversity of SIP experimental configurations exceeds the capacity of the MISIP standard to account for every attribute relevant for interpreting a given SIP sequence dataset. Several common, but poorly constrained, experimental attributes were not included in MISIP. For example, the standard does not account for the frequency and timing of multiple doses of isotopolog in a pulse-chase type of experiment, the spatial heterogeneity of isotopolog within the system during an incubation, whether the incubation took place *in situ*, or whether an experimental system was open or closed to the environment. Each of these experimental configurations may alter isotopic labeling due to the spatial-temporal variation in community metabolism and the influx or efflux of unlabeled or isotopically enriched isotopologs. MISIP can capture the kinetics of isotope enrichment of nucleic acids using the fields *nucleobase_atom_frac* and *isotopolog_incu_time*. Yet, measures of nucleic acid enrichment do not necessarily capture the full kinetics of isotope assimilation and cross-feeding, which may be captured in activity measurements, such as respiration or the isotopic labeling of other metabolites [34].

To address these limitations, we advise depositors to use the MISIP standard to guide data collection in planning and performing a SIP experiment. Once a user has completed the MISIP MIxS checklist, we recommend they provide any additional methodological information in the description of their methods referenced by the *sip_method* field. This will help ensure the relevant metadata not captured by the checklist are included in methodological descriptions available elsewhere in the public record. In addition to the provided metadata fields, if additional metadata or expanded information is needed to completely describe the experimental design (e.g., incubation parameters), details can be added using the *misc_parameter* term [MIXS:0000752]. We encourage users of the MISIP standard and community members that are domain experts to contribute to MISIP development by suggesting

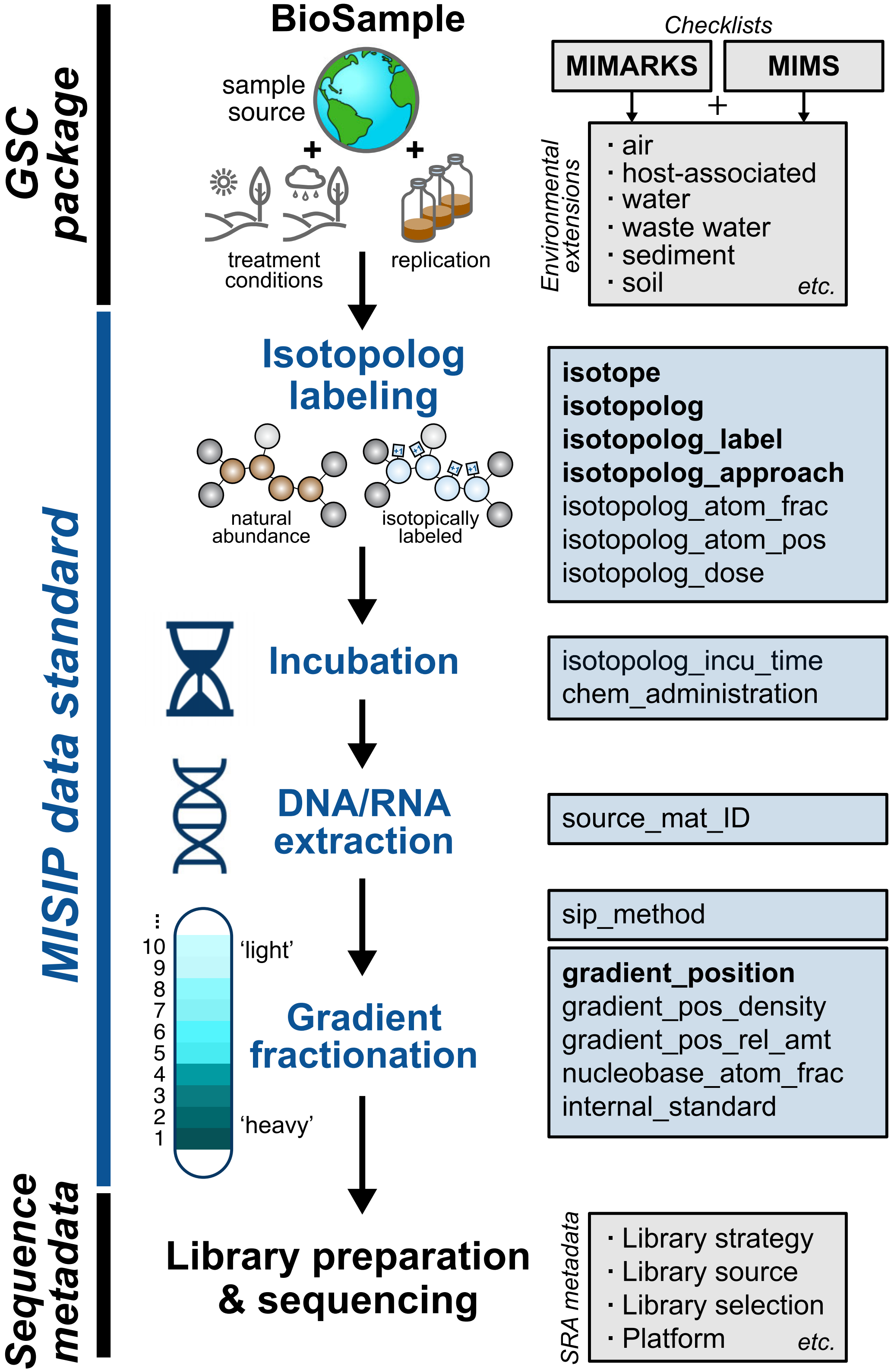

Figure 6: The MISIP standard captures metadata at several stages during a SIP experiment. This schematic provides an overview of the typical workflow of DNA/RNA-SIP experiment and the SIP-specific metadata (highlighted in blue boxes) that can be collected at each step. The related upstream and downstream metadata required for sequence data archival covered by other data standards are highlighted in black boxes.

Table 2: The MISIP standard uses a standardized vocabulary for instances where an expected value cannot be provided. There are 2 scenarios when this is necessary: (i) when users include a recommended field that includes samples with missing data and (ii) for unfractionated reference samples or other controls. The MISIP missing values vocabulary follows the latest standards issued.

Slot	Case 1: Data not collected	Case 2: Unfractionated reference samples	Criteria
isotope	Required	Not applicable—control sample	–
isotopolog	Required	None	–
isotopolog_label	Required	Not applicable—control sample	–
isotopolog_approach	Required	Not applicable—control sample	–
gradient_position	Required	−1	–
gradient_pos_density	Missing—not collected	Not applicable—control sample	Recommended
gradient_pos_rel_amt	Missing—not collected	Not applicable—control sample	Recommended
sip_method	Missing—not described	Missing—not described	Recommended
source_mat_id	Missing—not described	Missing—not described	Recommended
isotopolog_atom_frac	Missing—not collected	Not applicable—control sample	Recommended
isotopolog_atom_pos	Missing—not described	Not applicable—control sample	Recommended
isotopolog_dose	Missing—not described	Not applicable—control sample	Recommended
nucleobase_atom_frac	Missing—not collected	Not applicable—control sample	Recommended
isotopolog_incu_time	Not applicable	Not applicable—control sample	Recommended
hetero_isotopolog	Not applicable	Not applicable—control sample	Recommended
internal_standard	Missing—not collected	Not applicable—control sample	Recommended

additional terms, adding clarification, and providing feedback using the GSC issue tracker on the GSC GitHub repository [35]. This ensures that the standard meets the needs of the community and continues to evolve with newer methods.

The fields described in the MISIP standard are designed to be combined with at least one other MIxS checklist (MIMS or MIMARKS) plus a MIxS environmental extension. Users can also utilize controlled vocabulary from other existing MIxS extensions, such as the extensions for agricultural production systems [36] or for built environments [37]. The use of additional MIxS fields is recommended, such as nucleic acid extraction methods (*nucl_acid_ext*: MIXS:0000037), the primers used to generate gene marker data (*pcr_primers*: MIXS:0000046), and other contextual information in an environmental extension, such as the location sampled (*lat_lon*: MIXS:0000009) and chemical descriptions of the environment (e.g., *carb_nitro_ratio*: MIXS:0000310). Furthermore, the MISIP standard accommodates missing values using the vocabulary defined by the International Nucleotide Database Collaboration. In cases where a field that includes samples with missing data or when unfractionated reference samples are included, the user should provide the corresponding missing values (consult Table 2 for guidance). The use of missing values is typically necessary when compiling data from multiple SIP studies, as seen in Supplementary Table S10, where we have compiled Supplementary Tables S2 to S9 in machine-readable format as an example.

MISIP for qSIP

qSIP can be used to estimate the degree of isotopic enrichment of nucleic acids (or EAF) by resolving shifts in the buoyant density of individual nucleic acid sequences (Fig. 1B) [3]. The calculations necessary to perform qSIP require precise measurements of the fraction densities (*gradient_pos_density*) and relative DNA amount (*gradient_pos_rel_amt*). Thus, to perform qSIP, these 2 metadata fields are required rather than recommended. Practitioners of qSIP must determine the *gradient_pos_rel_amt* based on their sequencing approach by, for example, obtaining either the total gene abundance (in the case of amplicon data) or the total nucleic acid mass (for shotgun metagenomics) for the total amount of unfractionated nucleic acid sample added to the density gradient pool (Fig. 4). Other fields provide critical information for linking the quantitative enrichment of nucleic acids with environmental processes, including *isotopolog_incu_time* (for estimating process rates) and *isotopolog_dose*, *nucleobase_atom_frac*, and *isotopolog_atom_frac* (for estimating mass transfer). The ability to calculate rates of isotope incorporation with qSIP has been used to great effect in measuring the taxon-specific growth and mortality of soil populations [30], estimating microbial predator–prey dynamics [9], and disentangling the relationships between microbial communities, fluctuating environmental conditions, and biogeochemical processes [31].

Advice on the reuse of SIP sequence data

The MISIP standard is designed to encourage the reuse of SIP sequence data. To that end, we offer brief advice to assist in reanalyzing existing datasets. The primary focus of the analysis should be to contrast sequence data from equivalent gradient positions between paired samples supplied with either isotopically labeled or natural abundance isotopolog. We advise against making comparisons within a density gradient using isotopically labeled sequence data, except when a suitable standard is utilized and adjustments are made for the natural variation in GC content. Compositional differences due to GC content are easily controlled by comparison to sequence data from the paired natural abundance sample. That said, the inclusion of paired natural abundance samples is not required by MISIP, and in cases where it is missing, it may be possible to model the theoretical buoyant density distribution of natural abundance nucleic acid fragments from sequencing data [12, 38].

Although the MISIP standard aims to reduce error from mismatched gradient fractions, there is always a possibility of human error. For this reason, the inclusion of buoyant density information is highly recommended. However, in cases where density information is not provided, one should perform an iterative analysis using a “sliding window approach” to characterize differential abundance of nucleic acids [4], where adjacent gradient positions are combined to offset small variations in density at any one position.

Conclusions

For over two decades, the global SIP community has worked together to establish fundamental methods to connect microbial processes to nucleic acid sequencing and leverage the -omics analytical toolkit. Tracking the fate of an isotopically enriched isotopolog offers a unifying signal from which to integrate -omics data types (i.e., genomics, proteomic, metabolomic, etc.). While this approach offers a systems-level view of biological processes, it also requires adaptable frameworks to accommodate different SIP data types and metadata. The MISIP standard was developed to provide a foundation for these efforts by formalizing the minimum metadata requirements for any SIP-derived nucleic acid sequence and to formalize a common vocabulary for SIP metadata. By providing a shared vocabulary to guide metadata entry and validation, MISIP can assist in the development of bioinformatic software for SIP data analysis and the design of SIP data intake on platforms that extend a FAIR framework [13] to downstream data analysis, ensuring SIP continues to deepen our understanding of microbial communities in their environment.

The participation and support provided by the GSC governance and technical infrastructure is vital for standards development and for facilitating collaboration within scientific communities to develop new checklists and environmental extensions. To further the mission to increase the usability, reproducibility, maintainability, and consistency of archived sequencing data, the GSC has switched to managing MIxS standards via an open GitHub repository [35] and using LinkML tooling [39] with the release of MIxS version 6. The MISIP checklist will be included as part of the upcoming MIxS version 7 release (Fall 2024) and subsequently integrated into the International Nucleotide Sequence Database Collaboration and other databases. The entire suite of MIxS standards is updated regularly through approximately yearly releases. Once MISIP is released through MIxS, MISIP terms will be assigned permanent, resolvable, globally unique identifiers, and the latest stable authoritative version of the standard will always be available [35].

Additional Files

Supplementary Figure S1. An overview of the quality of SIP metadata archived at the Sequence Read Archive (SRA) from various environments. SIP studies accessioned in PubMed with a SRA containing more than five samples were evaluated for their metadata quality. SRA accessions which reported the isotopolog, isotopolog label status, and gradient position met the minimum ("Min") requirements, while those that reported at least one of these items were "Insufficient," and those that reported none were categorized as "None."

Supplementary Table S1. A list of all SIP experiments published on the SRA with ≥5 sequenced samples, including information on whether basic SIP metadata were available.

Supplementary Table S2. Example MISIP information for samples from Bradford et al., 2018 (PubMedID: 30483229). RNA-SIP was used to investigate the effect of toluene on the transcriptome of a BTEX (benzene, toluene, ethylbenzene, xylene)–contaminated aquifer. Note: the "Associated accession" is not required for MISIP and will be generated during the archival process. We have included it here to reference the original archived sequence data used in this example.

Supplementary Table S3. Example MISIP information for select samples from Barnett et al., 2021 (PubMedID: 34799453). The use of diverse carbon sources in agricultural soils was investigated using amplicon DNA-SIP. Note: the "Associated accession" is not required for MISIP and will be generated during the archival process. We have included it here to reference the original archived sequence data used in this example.

Supplementary Table S4. Example MISIP information for select samples from Papp et al., 2018 (PubMedID: 29439990). The relationship between growth rate and metabolic activity in soil microorganisms was investigated using qSIP. Note: the "Associated accession" is not required for MISIP and will be generated during the archival process. We have included it here to reference the original archived sequence data used in this example.

Supplementary Table S5. Example MISIP information for DNA-SIP libraries generated by Conover et al., 2021 (PubMedID: 34519133) using trimethylamine as the isotopolog. The aim was to test whether N from TMA was incorporated directly or secondarily via cross-feeding. Note: the "Associated accession" is not required for MISIP and will be generated during the archival process. We have included it here to reference the original archived sequence data used in this example.

Supplementary Table S6. Example MISIP information for select samples from Kong et al., 2020 (PubMedID: 31953339). The effect of manure on microbial use of rice residues in agricultural soil was investigated using amplicon DNA-SIP. Note: the "Associated accession" is not required for MISIP and will be generated during the archival process. We have included it here to reference the original archived sequence data used in this example.

Supplementary Table S7. Example MISIP information for samples from Thomas et al., 2021 (PubMedID: 33953365). The use of alginate (a macroalgal polysaccharide) by marine microbes was investigated using amplicon DNA-SIP. Note: the "Associated accession" is not required for MISIP and will be generated during the archival process. We have included it here to reference the original archived sequence data used in this example.

Supplementary Table S8. Example MISIP information for samples from Ding et al., 2014 (PubMedID: 25171335). The use of acetate by paddy soil microorganisms under different iron amendment conditions was investigated by amplicon RNA-SIP. Note: the "Associated accession" is not required for MISIP and will be generated during the archival process. We have included it here to reference the original archived sequence data used in this example.

Supplementary Table S9. Example MISIP information for samples from Macey et al., 2020 (PubMedID: 32156318). Methylotrophs in bulk soil and pea and wheat rhizosphere were characterized by amplicon DNA-SIP. Note: the "Associated accession" is not required for MISIP and will be generated during the archival process. We have included it here to reference the original archived sequence data used in this example.

Supplementary Table S10. A machine-readable version of all study data used to demonstrate curation according to the MISIP standard.

Abbreviations

CID: PubChem compound identification number; DOE: Department of Energy; DOI: digital object identifier; EAF: excess atom fraction; FAIR: Principles of "Findability, Accessibility, Interoperability, and Reusability"; GC: guanine-cytosine; GSC: Genomic Standards Consortium; IGSN: International Generic Sample Number; InChI: International Chemical Identifier; MIMARKS: Minimum Information about a MARKer gene Sequence; MIMS: Minimum Information about a Metagenome Sequence; MISIP: Minimum Information for any Stable Isotope Probing Sequence; MIxS: Minimum Information for any (x) Sequence; ORCID: Open Researcher and Contributor ID; qSIP: quantitative stable isotope probing; SIP: stable isotope probing.

Author Contributions

AS: Conceptualization, Methodology, Data Curation, Writing (Original Draft), and Writing (Review & Editing). **EWC**: Conceptualization, Methodology, Writing (Review & Editing), and Project Administration. **MS**: Conceptualization, Methodology, Writing (Review & Editing) and Project Administration. **BJK**: Conceptualization, Methodology, Writing (Review & Editing). **KB**: Conceptualization, Methodology, Writing (Review & Editing). **JAK**: Validation, Writing (Review & Editing). **MK**: Validation, Writing (Review & Editing). **CIH**: Data Curation, Writing (Review & Editing). **RLW**: Writing (Review & Editing). **LMS**: Writing (Review & Editing). **RCW**: Conceptualization, Methodology, Data Curation, Writing (Original Draft), and Writing (Review & Editing), and Project Administration

Acknowledgments

We thank Kaitlyn Johnson and Payton Taylor for their help curating the list of SIP studies and Drs. Samuel Barnett, Ning Ling, and Rex Malmstrom for providing valuable feedback on the MISIP standard and manuscript. We are grateful for the guidance received from members of the GSC's CIG and board members, specifically Mark Miller for the implementation of LinkML in the GSC.

Funding

Partial funding for RCW was provided by the USDA National Institute of Food and Agricultural (https://ror.org/03wjh4t84) Hatch Grant program (IND90024429), administered by Purdue University (https://ror.org/05p8z3f47). Support for the work conducted by EWC on behalf of the Systems Biology Knowledgebase (see for KBase: https://ror.org/01znn6x10) was provided by the Genomic Science Program in the US Department of Energy (DOE) Office of Science's Biological and Environmental Research (BER) Program under contract DE-AC02-05CH11231, administered by the Lawrence Berkeley National Lab (https://ror.org/02jbv0t02). Support for the work conducted by MLS on behalf of the National Microbiome Data Collaborative (https://ror.org/05cwx3318) was also provided by the DOE-BER's Genomic Science Program under contract DE-AC05-76RL01830, administered by the Pacific Northwest National Lab (https://ror.org/05h992307). Support for the work conducted by M.K. at the Joint Genome Institute (https://ror.org/04xm1d337) was provided by the DOE Office of Science under contract DE-AC02-05CH11231. Support for the work by B.K., at Northern Arizona University (https://ror.org/0272j5188), and J.A.K., at Lawrence Livermore National Lab (https://ror.org/03z21mf61), was provided by the DOE-BER's Genomic Science Program "Microbes Persist" Soil Microbiome Scientific Focus Area (award #SCW1632) under contract DE-AC52-07NA27344.

Data Availability

The data supporting this article, including a tab-separated, machine-readable version of a compilation of SIP studies in the MISIP format (Supplementary Table S10), are available in the *GigaScience* repository, GigaDB [40]. Details of the 138 SIP studies utilized for evaluating the quality of metadata curation in public archives are presented in Supplementary Table S1. The MISIP checklist will be included as part of the upcoming GSC-MIxS version 7 release (Fall 2024) available through the GSC [35] and subsequently integrated into the International Nucleotide Sequence Database Collaboration archives (EBI, NCBI, DDBJ) and other databases.

Competing Interests

The authors declare that they have no competing interests.

References

1. Radajewski S, Ineson P, Parekh NR, et al. Stable-isotope probing as a tool in microbial ecology. Nature 2000;403:646–49. https://doi.org/10.1038/35001054.
2. Manefield M, Whiteley AS, Griffiths RI, et al. RNA stable isotope probing, a novel means of linking microbial community function to phylogeny. Appl Environ Microb 2002;68:5367–73. https://doi.org/10.1128/AEM.68.11.5367-5373.2002.
3. Hungate BA, Mau RL, Schwartz E, et al. Quantitative microbial ecology through stable isotope probing. Appl Environ Microb 2015;81:7570–81. https://doi.org/10.1128/AEM.02280-15.
4. Youngblut ND, Barnett SE, Buckley DH. HTSSIP: an R package for analysis of high throughput sequencing data from nucleic acid stable isotope probing (SIP) experiments. PLoS One 2018;13:e0189616. https://doi.org/10.1371/journal.pone.0189616.
5. Vyshenska D, Sampara P, Singh K, et al. A standardized quantitative analysis strategy for stable isotope probing metagenomics. mSystems 2023;8:e01280–22. https://doi.org/10.1128/msystems.01280-22.
6. Nuccio EE, Blazewicz SJ, Lafler M, et al. HT-SIP: a semi-automated stable isotope probing pipeline identifies cross-kingdom interactions in the hyphosphere of arbuscular mycorrhizal fungi. Microbiome 2022;10:199. https://doi.org/10.1186/s40168-022-01391-z.
7. Wilhelm RC, Singh R, Eltis LD, et al. Bacterial contributions to delignification and lignocellulose degradation in forest soils with metagenomic and quantitative stable isotope probing. ISME J 2019;13:413–29. https://doi.org/10.1038/s41396-018-0279-6.
8. Barnett SE, Youngblut ND, Koechli CN, et al. Multisubstrate DNA stable isotope probing reveals guild structure of bacteria that mediate soil carbon cycling. Proc Natl Acad Sci U S A 2021;118(47):e2115292118. https://doi.org/10.1073/pnas.2115292118.
9. Hungate BA, Marks JC, Power ME, et al. The functional significance of bacterial predators. mBio 2021;12:e00466–21. https://doi.org/10.1128/mBio.00466-21.
10. Neufeld JD, Vohra J, Dumont MG, et al. DNA stable-isotope probing. Nat Protoc 2007;2:860–66. https://doi.org/10.1038/nprot.2007.109.
11. Youngblut ND, Buckley DH. Intra-genomic variation in G+C content and its implications for DNA stable isotope probing. Environ Microbiol Rep 2014;6:767–75. https://doi.org/10.1111/1758-2229.12201.
12. Wilhelm RC, Pepe-Ranney C, Weisenhorn P, et al. Competitive exclusion and metabolic dependency among microorganisms structure the cellulose economy of an agricultural soil. mBio 2021;12:e03099–20. https://doi.org/10.1128/mBio.03099-20.
13. Wilkinson MD, Dumontier M, IjJ A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.

14. Yilmaz P, Kottmann R, Field D, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 2011;29:415–20. https://doi.org/10.1038/nbt.1823.
15. Kottmann R, Gray T, Murphy S; et al.; Genomic Standards Consortium A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). OMICS 2008;12(2):115–21.
16. Field D, Amaral-Zettler L, Cochrane G, et al. The Genomic Standards Consortium. PLoS Biol 2011;9:e1001088. https://doi.org/10.1371/journal.pbio.1001088.
17. Simpson A, Charlson EMW, Smith M, et al. A data standard for the reuse and reproducibility of any stable isotope probing-derived nucleic acid sequence (MISIP). BioRxiv 2023. v1. https://doi.org/10.1101/2023.07.13.548835
18. Coplen TB. Guidelines and recommended terms for expression of stable-isotope-ratio and gas-ratio measurement results. Rapid Comm Mass Spectrometry 2011;25:2538–60. https://doi.org/10.1002/rcm.5129.
19. Morrissey EM, Mau RL, Schwartz E, et al. Phylogenetic organization of bacterial activity. ISME J 2016;10:2336–40. https://doi.org/10.1038/ismej.2016.28.
20. Wilhelm RC, DeRito CM, Shapleigh JP, et al. Phenolic acid-degrading paraburkholderia prime decomposition in forest soil. ISME Commun 2021;1:4. https://doi.org/10.1038/s43705-021-00009-z.
21. Mau RL, Liu CM, Aziz M, et al. Linking soil bacterial biodiversity and soil carbon stability. ISME J 2015;9:1477–80. https://doi.org/10.1038/ismej.2014.205.
22. Wilhelm R, Szeitz A, Klassen TL, et al. Efficient quantitation of 13C-enriched nucleic acids via ultrahigh-performance liquid chromatography-tandem mass spectrometry for applications in stable isotope probing. Appl Environ Microb 2014;80:7206–11. https://doi.org/10.1128/AEM.02223-14.
23. Pepe-Ranney C, Campbell AN, Koechli CN, et al. Unearthing the ecology of soil microorganisms using a high resolution DNA-SIP approach to explore cellulose and xylose metabolism in soil. Front Microbiol 2016;7:703. https://doi.org/10.3389/fmicb.2016.00703.
24. Thomas F, Le Duff N, Wu T-D, et al. Isotopic tracing reveals single-cell assimilation of a macroalgal polysaccharide by a few marine flavobacteria and gammaproteobacteria. ISME J 2021;15:3062–75. https://doi.org/10.1038/s41396-021-00987-x.
25. Teytelman L, Stoliartchouk A, Kindler L, et al. Protocols.Io: virtual communities for protocol development and discussion. PLoS Biol 2016;14:e1002538. https://doi.org/10.1371/journal.pbio.1002538.
26. Damerow JE, Varadharajan C, Boye K, et al. Sample identifiers and metadata to support data management and reuse in multidisciplinary ecosystem sciences. Data Sci J 2021;20:11. https://doi.org/10.5334/dsj-2021-011.
27. Dijkstra P, Dalder JJ, Selmants PC, et al. Modeling soil metabolic processes using isotopologue pairs of position-specific 13C-labeled glucose and pyruvate. Soil Biol Biochem 2011;43:1848–57. https://doi.org/10.1016/j.soilbio.2011.05.001.
28. Heller SR, McNaught A, Pletnev I, et al. InChI, the IUPAC International Chemical Identifier. J Cheminform 2015;7:23. https://doi.org/10.1186/s13321-015-0068-4.
29. Wilhelm RC. Guide to generating international chemical identifier (InChI) label for isotopically labelled compounds V.1. Protocols.io 2023:v1. dx.doi.org/10.17504/protocols.io.rm7vzx9nrgx1/v1
30. Koch BJ, McHugh TA, Hayer M, et al. Estimating taxon-specific population dynamics in diverse microbial communities. Ecosphere 2018;9(1):e02090. https://doi.org/10.1002/ecs2.2090.
31. Blazewicz SJ, Hungate BA, Koch BJ, et al. Taxon-specific microbial growth and mortality patterns reveal distinct temporal population responses to rewetting in a California grassland soil. ISME J 2020;14:1520–32. https://doi.org/10.1038/s41396-020-0617-3.
32. Niu J, Kasuga I, Kurisu F, et al. Evaluation of autotrophic growth of ammonia-oxidizers associated with granular activated carbon used for drinking water purification by DNA-stable isotope probing. Water Res 2013;47:7053–65. https://doi.org/10.1016/j.watres.2013.07.056.
33. Xia W-W, Zhao J, Zheng Y, et al. Active soil nitrifying communities revealed by in situ transcriptomics and microcosm-based stable-isotope probing. Appl Environ Microb 2020;86(23):e01807–20. https://doi.org/10.1128/AEM.01807-20.
34. Wilhelm RC, Barnett SE, Swenson TL, et al. Tracing carbon metabolism with stable isotope metabolomics reveals the legacy of diverse carbon sources in soil. Appl Environ Microb 2022;88(22):e00839–22. https://doi.org/10.1128/aem.00839-22.
35. Genomic Standards Consortium. MIxS: minimum information about any (x) sequence specifications. https://github.com/GenomicsStandardsConsortium/mixs. Accessed 13 December 2023.
36. Dundore-Arias JP, Eloe-Fadrosh EA, Schriml LM, et al. Community-driven metadata standards for agricultural microbiome research. Phytobiomes J 2020;4:115–21. https://doi.org/10.1094/PBIOMES-09-19-0051-P.
37. Glass EM, Dribinsky Y, Yilmaz P, et al. MIxS-BE: a MIxS extension defining a minimum information standard for sequence data from the built environment. ISME J 2014;8:1–3. https://doi.org/10.1038/ismej.2013.176.
38. Barnett SE, Buckley DH. Simulating metagenomic stable isotope probing datasets with MetaSIPSim. BMC Bioinf 2020;21:37. https://doi.org/10.1186/s12859-020-3372-6.
39. Moxon S, Solbrig H, Unni D, et al. The Linked Data Modeling language (LinkML): 2021 International Conference on Biomedical Ontologies, ICBO 2021. In: CEUR Workshop Proceedings. CEUR-WS.org, 148–51. 3073.
40. Simpson AC, Wood-Charlson EM, Smith M, et al. Supporting data for “MISIP: A Data Standard for the Reuse and Reproducibility of Any Stable Isotope Probing-Derived Nucleic Acid Sequence and Experiment.” GigaScience Database. 2024. https://doi.org/10.5524/102569.(?PMU?)

Received: June 1, 2024. **Revised:** July 29, 2024. **Accepted:** August 27, 2024